Procesamiento de Información Temporal

Tema 2:
Modelos Ocultos de Markov (HMM)
Parte 2: Formalismo y Algorítmica de HMM

Alicia Lozano Díez
alicia.lozano@uam.es
Audias – Audio, Data Intelligence and Speech
Universidad Autónoma de Madrid
http://audias.ii.uam.es
Materiales basados en los de Daniel Ramos Castro

1

Modelos Ocultos de Markov (HMM): Formalismo y Algorítmica

2

Definición formal de un HMM

- Un HMM queda caracterizado por los siguientes elementos:
  - 1. El número de estados en el modelo, N.
    - Denotamos cada estado como S_i
    - Denotamos el estado en el instante t como q_t
    - Si el sistema está en el estado S_i en el instante t escribimos $q_t=S_i$
  - 2. El número de símbolos observables, M
    - Denotamos a cada símbolo observable como v_j
    - Denotamos la observación en el instante t como O_t
    - Si la observación en el instante t es v_j escribimos $O_t=v_j$
  - 3. La matriz de probabilidades de transición

$$A = \{a_{ij}\} \qquad a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \qquad 1 \leq i, j \leq N.$$

Definición formal de un HMM (ii)

- Un HMM queda caracterizado por los siguientes elementos (ii):
  - 4. La distribución de probabilidad de observación en cada estado j (*likelihood*)

$$B = \{b_j(k)\}, \quad b_j(k) = P[v_k \text{ at } t | q_t = S_j], \qquad 1 \leq j \leq N$$

$$1 \leq k \leq M.$$

  - 5. La probabilidad inicial de ocupación de cada estado

$$\pi = \{\pi_i\} \qquad \pi_i = P[q_1 = S_i], \qquad 1 \leq i \leq N.$$

- Por conveniencia denotamos todos los parámetros del HMM (incluyendo N y M) como: $\lambda = (A, B, \pi)$

HMMs como modelos generativos

- Muchas veces se dice que los HMMs son modelos generativos
- Esto es porque un HMM nos define un procedimiento muy sencillo de generar observaciones compatibles con el modelo
- Para generar una secuencia de T observaciones compatibles con un HMM se puede emplear el siguiente procedimiento:
  - 1. Elegir un estado inicial $q_1=S_i$ de acuerdo con la distribución de probabilidad inicial de estados, π.
  - 2. Hacer t=1.
  - 3. Elegir $O_t=v_k$ de acuerdo con la distribución de probabilidad de observación en el estado S_i, $b_i(k)$.
  - 4. Pasar a un nuevo estado $q_t+1=S_j$ de acuerdo con las probabilidades de transición desde el estado S_i, a_{ij}.
  - 5. Hacer t = t+1 y volver al punto 3 si t<T, en caso contrario terminar.

Los tres problemas básicos de los HMMs

- Hay tres problemas básicos que se deben resolver (y que afortunadamente están resueltos) para que los HMMs sean útiles en aplicaciones prácticas
- **Problema 1: Problema de puntuación:** Dada una secuencia de observaciones y un modelo, ¿cómo calcular la probabilidad de observar la secuencia vista dado el modelo (*likelihood*)?
- **Problema 2: Problema de reconocimiento de estados:** Dada una secuencia de observaciones y un modelo, ¿cuál es la secuencia de estados que mejor “explica” las observaciones?
- **Problema 3: Problema de entrenamiento:** Dado un conjunto de observaciones de entrenamiento ¿Cómo ajustamos los parámetros del modelo para maximizar la probabilidad de observar el conjunto de entrenamiento dado el modelo (*maximum likelihood*)?

Solución al Problema 1: Solución directa

- Tenemos un HMM: $\lambda = (A, B, \pi)$
- Queremos calcular: $P(O|\lambda)$ $O = O_1 O_2 \cdots O_T$
- Si suponemos que la secuencia de estados es: $Q = q_1 \, q_2 \cdots q_T$
- Entonces, asumiendo que las observaciones son independientes estadísticamente dada la secuencia de estados y el modelo, tenemos:

$$P(O|Q, \lambda) = \prod_{t=1}^{T} P(O_t|q_t, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdots b_{q_T}(O_T).$$

- Por otro lado, la probabilidad conjunta de O y Q es:

$$P(O, Q|\lambda) = P(O|Q, \lambda) \, P(Q|\lambda).$$

- Donde $P(Q|\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$.
  - Por la propiedad de Markov (un estado solo depende de anterior en el tiempo)

Solución al Problema 1: Solución directa (ii)

- La probabilidad de O dado el modelo se obtiene sumando las anteriores probabilidades para todos los posibles caminos (secuencias de estados ocultos):
  - Marginalización

$$P(O|\lambda) = \sum_{\text{all } Q} P(O|Q, \lambda) \, P(Q|\lambda)$$

$$= \sum_{q_1, q_2, \cdots, q_T} \pi_{q_1} b_{q_1}(O_1) \, a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(Q_T).$$

- Problema del cálculo directo: Requiere $O(2TN^T)$ operaciones
- Para N=5 estados, T=100 observaciones → 10^{72} operaciones
- No escalable, no tratable

Solución al Problema 1: Algoritmo Forward / Backward (i)

- Consiste en definir la variable forward: $\alpha_t(i) = P(O_1 O_2 \cdots O_t, q_t = S_i|\lambda)$
  - Representa la probabilidad de observar la secuencia parcial $O_1, \ldots, O_t$ hasta el instante t y estar en el estado S_i en dicho instante t.
- La variable forward se puede calcular con el siguiente algoritmo:

1) Initialization:

$$\alpha_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N.$$

2) Induction:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}), \qquad 1 \le t \le T-1$$
$$1 \le j \le N.$$

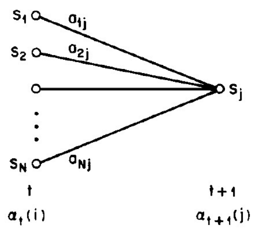
3) Termination:

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

Solución al Problema 1: Algoritmo Forward / Backward (ii)

- El punto clave es el paso de inducción, que permite calcular las variables forward en el instante t+1 a partir de las variables forward en el t, de las probabilidades de transición y las de observación:
- Para cada estado en el instante t+1 tenemos que hacer las siguientes operaciones:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1})$$

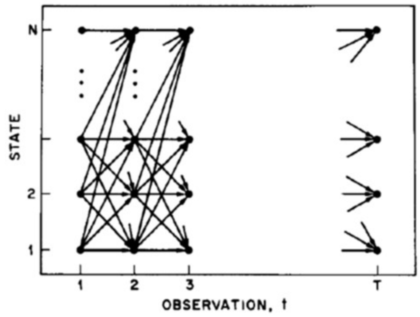


- En total N+1 multiplicaciones y N-1 sumas

[Rabiner89]

Solución al Problema 1: Algoritmo Forward / Backward (iii)

- Para T instantes de tiempo y N estados debemos repetir esa operación básica NxT veces



[Rabiner89]

- El coste computacional total es del orden de TN^2 en lugar de $2TN^T$
  - Para N=5, T=100 → 3000 operaciones en lugar de 10^{72}

Solución al Problema 1: Algoritmo Forward / Backward (iv)

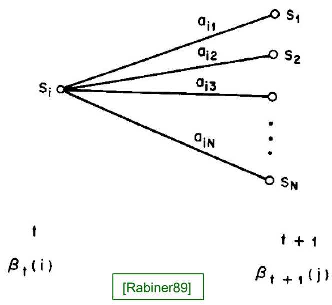
- Lo visto hasta ahora es el algoritmo forward, que permite por sí solo calcular la probabilidad de la observación dado el modelo
- Alternativamente, se puede usar la variable backward:

$$\beta_t(i) = P(O_{t+1}\, O_{t+2} \cdots O_T | q_t = S_i, \lambda)$$

  - Representa la probabilidad de observar la secuencia parcial O_{t+1}, …, O_T desde el instante t+1 y estar en el estado S_i en el instante t.
- En realidad para calcular la probabilidad de la observación dado el modelo se puede emplear o el algoritmo forward o el algoritmo backward
  - Sólo es necesario uno para resolver este problema
  - Pero utilizando los dos podremos resolver más fácilmente otros problemas
- Para “backward”: necesaria la secuencia de observaciones completa
  - No es necesario en “forward”, que podría hacerse de forma secuencial

Solución al Problema 1: Algoritmo Forward / Backward (v)

- El algoritmo backward



[Rabiner89]

1) Initialization:

$$\beta_T(i) = 1, \qquad 1 \leq i \leq N.$$

2) Induction:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j),$$

$$t = T-1, T-2, \cdots, 1, \; 1 \leq i \leq N.$$

- El coste computacional es idéntico al forward

Los tres problemas básicos de los HMMs

- Hay tres problemas básicos que se deben resolver (y que afortunadamente están resueltos) para que los HMMs sean útiles en aplicaciones prácticas
- **Problema 1: Problema de puntuación:** Dada una secuencia de observaciones y un modelo, ¿cómo calcular la probabilidad de observar la secuencia vista dado el modelo?
- **Problema 2: Problema de reconocimiento de estados: Dada una secuencia de observaciones y un modelo, ¿cuál es la secuencia de estados que mejor "explica" las observaciones?**
- **Problema 3: Problema de entrenamiento:** Dado un conjunto de observaciones de entrenamiento ¿Cómo ajustamos los parámetros del modelo para maximizar la probabilidad de observar el conjunto de entrenamiento dado el modelo?

Solución al Problema 2: Criterios para elegir la secuencia de modelos "óptima"

- Posibles criterios:
  - Elegir en cada instante de tiempo el estado más probable
    - Consiste en maximizar en cada instante de tiempo la variable:

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

    - Problema: Puede ocurrir que la secuencia de estados (por ejemplo una secuencia de fonemas) no tenga sentido
  - Elegir el camino completo de estados con mayor probabilidad total
    - Consiste en encontrar la secuencia de estados que maximiza globalmente

$$P(O|Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdots b_{q_T}(O_T).$$

    - Este es un problema que tiene más sentido que el anterior
    - En la analogía con los fonemas reconocidos consistiría en encontrar la secuencia de estados (fonemas) válida más probable
    - Es el problema a resolver en reconocimiento de voz
    - Se resuelve con el algoritmo de Viterbi

Solución al Problema 2: Estado más probable en cada instante de tiempo

- Para calcular el estado más probable en cada instante de tiempo debemos encontrar una forma eficiente de calcular la probabilidad de ocupación de cada estado en el instante t

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

- Una forma posible es calculando las variables forward y backward y combinarlas para calcular la variable anterior

$$\gamma_t(i) = \frac{\alpha_t(i)\ \beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\ \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\ \beta_t(i)}$$

- Ahora ya sólo queda tomar el estado más probable:

$$q_t = \underset{1 \le i \le N}{\operatorname{argmax}} [\gamma_t(i)], \qquad 1 \le t \le T.$$

Solución al Problema 2: Viterbi (i)

- Para encontrar la secuencia de estados (Q) más probable para una secuencia de observaciones (O) dada definimos la siguiente variable auxiliar:

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P[q_1 q_2 \cdots q_t = i, O_1 O_2 \cdots O_t | \lambda]$$

- Esta variable representa la mejor puntuación obtenida a través de una única secuencia de estados ($q_1, q_2, \ldots q_{t-1}$) hasta llegar, en el instante t, al estado i
- La clave del algoritmo Viterbi es que si conocemos estas variables auxiliares en el instante t para todos los estados, podemos calcularlas para el siguiente instante (t+1) y para todos los estados con:

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \cdot b_j(O_{t+1}).$$

- Si queremos obtener la secuencia de estados no nos basta con estas variables, tenemos que saber también el estado i que maximiza

Solución al Problema 2: Viterbi (ii)

- El algoritmo completo de Viterbi queda de la siguiente forma
- El backtracking consiste en volver desde el instante final al inicial viendo qué estados maximizaban cada paso de la recursión
- El algoritmo es casi idéntico al forward sustituyendo la suma de la recursión por una maximización

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N$$

$$\psi_1(i) = 0.$$

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T$$
$$1 \le j \le N$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \le t \le T$$
$$1 \le j \le N.$$

3) Termination:

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

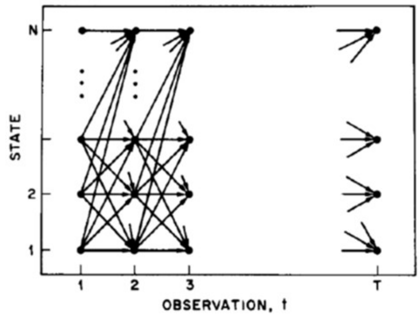
$$q_T^* = \operatorname*{argmax}_{1 \le i \le N} [\delta_T(i)].$$

4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1.$$

Solución al Problema 2: Viterbi (iii)

- La forma de operar con el algoritmo es también similar:



[Rabiner89]

- Y el número de operaciones es también $O(TN^2)$

Los tres problemas básicos de los HMMs

- Hay tres problemas básicos que se deben resolver (y que afortunadamente están resueltos) para que los HMMs sean útiles en aplicaciones prácticas
- **Problema 1: Problema de puntuación:** Dada una secuencia de observaciones y un modelo, ¿cómo calcular la probabilidad de observar la secuencia vista dado el modelo?
- **Problema 2: Problema de reconocimiento de estados:** Dada una secuencia de observaciones y un modelo, ¿cuál es la secuencia de estados que mejor "explica" las observaciones?
- **Problema 3: Problema de entrenamiento: Dado un conjunto de observaciones de entrenamiento ¿Cómo ajustamos los parámetros del modelo para maximizar la probabilidad de observar el conjunto de entrenamiento dado el modelo?**

Solución al Problema 3: Algoritmo de Baum-Welch (i)

- Formalmente el algoritmo Baum-Welch (y el EM) comienza con la definición de una función auxiliar:

$$Q(\lambda, \bar{\lambda}) = \sum_{Q} P(Q|O, \lambda) \log [P(O, Q|\bar{\lambda})]$$

- Esta función depende de los parámetros anteriores del modelo (λ) y de la nueva estimación de los parámetros: $\bar{\lambda}$.
- La teoría del algoritmo E-M nos dice que maximizar esta función auxiliar respecto a los nuevos parámetros nos lleva a una verosimilitud mayor:

$$\max_{\bar{\lambda}} [Q(\lambda, \bar{\lambda})] \Rightarrow P(O|\bar{\lambda}) \geq P(O|\lambda).$$

- Si utilizamos los nuevos parámetros en la función auxiliar como parámetros anteriores del modelo y repetimos el proceso varias veces seguiremos aumentando la verosimilitud hasta que el algoritmo converja o la variación de verosimilitud sea muy pequeña

Solución al Problema 3: Algoritmo de Baum-Welch: Expectation (ii)

- El primer paso (Expectation step) del algoritmo consiste en calcular las partes de la siguiente ecuación que dependen del modelo anterior

$$Q(\lambda, \bar{\lambda}) = \sum_{Q} P(Q|O, \lambda) \log [P(O, Q|\bar{\lambda})]$$

- En particular hay que calcular $P(Q|O, \lambda)$
- Es decir, las probabilidades de todas las secuencias de estados dados el modelo anterior y las observaciones
- En realidad basta con estimar (dados el modelo anterior y las observaciones):
  - La probabilidad de estar en el estado S_i en el instante t
  - El número esperado de transiciones desde el estado S_i
  - El número esperado de transiciones desde el estado S_i al estado S_j

Solución al Problema 3: Algoritmo de Baum-Welch: Expectation (iii)

- Dados el modelo anterior y las observaciones:
  - La probabilidad de estar en el estado Si en el instante t viene dada por la probabilidad de ocupación del estado que definíamos anteriormente y se podía calcular en función de las variables forward y backward

$$\gamma_t(i) = P(q_t = S_i | O, \lambda)$$

$$\gamma_t(i) = \frac{\alpha_t(i)\,\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\,\beta_t(i)}$$

  - El número esperado de transiciones desde el estado S_i se puede calcular en función de éstas como:

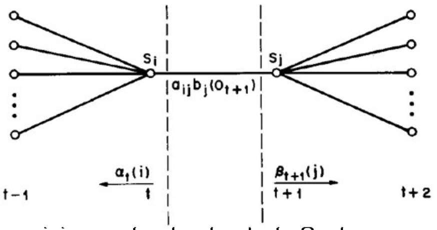
$$\sum_{t=1}^{T-1} \gamma_t(i)$$

Solución al Problema 3: Algoritmo de Baum-Welch: Expectation (iv)

- Dados el modelo anterior y las observaciones:
  - Para calcular el número esperado de transiciones desde el estado S_i al estado S_j definimos primero la probabilidad de la transición desde el estado S_i al estado S_j

$$\xi_t(i, j) = P(q_t = S_i,\, q_{t+1} = S_j | O, \lambda).$$

$$= \frac{\alpha_t(i)\, a_{ij} b_j(O_{t+1})\, \beta_{t+1}(j)}{P(O|\lambda)}$$



  - Con esto el número esperado de transiciones desde el estado S_i al estado S_j es:

$$\sum_{t=1}^{T-1} \xi_t(i, j)$$

Solución al Problema 3: Algoritmo de Baum-Welch: Maximization (v)

- Una vez tenemos estimados el número medio de todas las transiciones y el número medio de veces en cada estado maximizamos la función $Q(\lambda, \bar{\lambda}) = \sum_{Q} P(Q|O, \lambda) \log [P(O, Q|\bar{\lambda})]$
- Respecto a los datos del nuevo modelo, obteniendo nuevas estimaciones del modelo dadas por las fórmulas

$$\bar{\pi}_i = \text{expected frequency (number of times) in state } S_i \text{ at time } (t = 1) = \gamma_1(i)$$

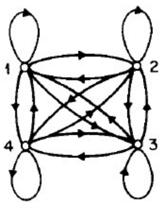
$$\bar{a}_{ij} = \frac{\text{expected number of transitions from state } S_i \text{ to state } S_j}{\text{expected number of transitions from state } S_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\bar{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } v_k}{\text{expected number of times in state } j} = \frac{\sum_{\substack{t=1 \\ \text{s.t. } O_t = v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$$

Topología de los HMMs

- Viene determinada por el número de estados y la interconexión entre ellos mediante probabilidades de transición no nulas
  - En definitiva, por el número de estados y la matriz de transición
  - Nota: Una probabilidad de transición nula inicialmente se mantiene siempre nula en el proceso de reestimación Baum-Welch
- HMMs ergódicos
  - Son modelos en los que la probabilidad de transición de un estado a cualquier otro es no nula

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

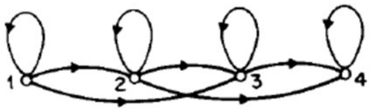


[Rabiner89]

Topología de los HMMs (ii)

- Modelos de Bakis o de izquierda a derecha
  - Son tales que aseguran que una vez hemos salido de un estado nunca podemos volver a él
  - Son apropiados para modelar señales que varían en el tiempo como la voz (una vez hemos terminado de pronunciar un fonema pasamos a otro pero no volvemos al mismo, en todo caso podemos pasar a otra realización del mismo fonema)

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ 0 & a_{22} & a_{23} & a_{24} \\ 0 & 0 & a_{33} & a_{34} \\ 0 & 0 & 0 & a_{44} \end{bmatrix}$$



[Rabiner89]

HMMs con funciones densidad de probabilidad de observación continuas

- Hasta ahora hemos asumido que las observaciones eran discretas
- En la práctica en reconocimiento de voz las observaciones son vectores de parámetros (MFCCs) que se consideran continuos
- Inicialmente estos vectores de parámetros se discretizaban mediante el proceso de cuantificación vectorial (VQ)
- Pero en la actualidad suelen funcionar modelando la probabilidad de las observaciones con una función densidad de probabilidad continua
- Habitualmente esta función densidad de probabilidad de la observación para cada estado se define como una mezcla de M Gaussianas multidimensionales:

$$b_j(\mathbf{O}) = \sum_{m=1}^{M} c_{jm} \mathfrak{N}[\mathbf{O}, \boldsymbol{\mu}_{jm}, \boldsymbol{U}_{jm}], \qquad 1 \leq j \leq N$$

HMMs con funciones densidad de probabilidad de observación continuas (ii)

- La utilización de fdps continuas complica especialmente el proceso de estimación del modelo (Baum-Welch), pues ahora hay que estimar los vectores de medias y las matrices de covarianzas para cada estado, así como el peso de cada Gaussiana.
- Para el paso de Estimación conviene definir la función auxiliar

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\,\beta_t(j)}{\sum_{j=1}^{N}\alpha_t(j)\,\beta_t(j)}\right]\left[\frac{c_{jk}\mathfrak{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jk}, \mathbf{U}_{jk})}{\sum_{m=1}^{M} c_{jm}\mathfrak{N}(\mathbf{O}_t, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm})}\right]$$

- Esta variable representa la probabilidad de estar en el estado j en el instante de tiempo t, con la Gaussiana k explicando la observación O_t.

HMMs con funciones densidad de probabilidad de observación continuas (iii)

- En función de esta variable auxiliar, en el paso de Maximización debemos reestimar los siguientes parámetros adicionales:

$$\bar{c}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{M}\gamma_t(j,k)} \qquad \bar{\boldsymbol{\mu}}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)\cdot\mathbf{O}_t}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

$$\overline{\mathbf{U}}_{jk} = \frac{\sum_{t=1}^{T}\gamma_t(j,k)\cdot(\mathbf{O}_t-\boldsymbol{\mu}_{jk})(\mathbf{O}_t-\boldsymbol{\mu}_{jk})'}{\sum_{t=1}^{T}\gamma_t(j,k)}$$

HMMs con funciones densidad de probabilidad de observación continuas (iv)

- En reconocimiento de voz se emplean principalmente HMMs:
  - Con topología Bakis
  - Con probabilidades de observación por estados definidas con fdps continuas
  - Con fdps continuas modeladas con mezclas de Gaussianas multidimensionales
- Habitualmente se emplean como parámetros para el reconocimiento de voz MFCCs
  - Se ha demostrado que las componentes de los MFCC obtenidos a partir de la voz están aproximadamente incorreladas entre ellas
  - Esto justifica la utilización de matrices de covarianza diagonales en las fdps.
  - Esto nos reduce mucho el número de parámetros a estimar y facilita el entrenamiento con un número de datos reducido.

Referencias Bibliográficas

- [Rabiner89] L. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77(2), 1989.
- [Benesti08] Benesti et al. "HMMs and Related Speech Recognition Technologies". In Handbook of Speech Processing, Chapter 27, Springer, 2008.